计算机科学
COMPUTER SCIENCE

基于遗憾探索的竞争网络强化学习智能推荐方法研究

洪志理，赖俊，曹雷，陈希亮，徐志雄

引用本文

洪志理，赖俊，曹雷，陈希亮，徐志雄. 基于遗憾探索的竞争网络强化学习智能推荐方法研究[J]. 计算机科学, 2022, 49(6): 149-157.

HONG Zhi-li, LAI Jun, CAO Lei, CHEN Xi-liang, XU Zhi-xiong. Study on Intelligent Recommendation Method of Dueling Network Reinforcement Learning Based on Regret Exploration[J]. Computer Science, 2022, 49(6): 149-157.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.210600226

基于遗憾探索的竞争网络强化学习智能推荐方法研究

洪志理　赖　俊　曹　雷　陈希亮　徐志雄
陆军工程大学指挥控制工程学院　南京 210007
(2206851664@qq.com)

摘　要　近年来,深度强化学习在推荐系统中的应用受到了越来越多的关注。在已有研究的基础上提出了一种新的推荐模型RP-Dueling,该模型在深度强化学习Dueling-DQN的基础上加入了遗憾探索机制,使算法根据训练程度自适应地动态调整"探索-利用"占比。该算法实现了在拥有大规模状态空间的推荐系统中捕捉用户动态兴趣和对动作空间的充分探索。在多个数据集上进行测试,所提算法在*MAE*和*RMSE*两个评价指标上的最优平均结果分别达到了0.16和0.43,比目前的最优研究结果分别降低了0.48和0.56,实验结果表明所提模型优于目前已有的传统推荐模型和基于深度强化学习的推荐模型。

关键词:推荐系统;深度强化学习;Dueling-DQN;RP-Dueling;动态兴趣;遗憾探索
中图法分类号　TP181

Study on Intelligent Recommendation Method of Dueling Network Reinforcement Learning Based on Regret Exploration

HONG Zhi-li,LAI Jun,CAO Lei,CHEN Xi-liang and XU Zhi-xiong
Command & Control Engineering College,Army Engineering University of PLA,Nanjing 210007,China

Abstract　In recent years,the application of deep reinforcement learning in recommendation system has attracted much attention. Based on the existing research,this paper proposes a new recommendation model RP-Dueling,which is based on the deep reinforcement learning Dueling-DQN algorithm,and adds the regret exploration mechanism to make the algorithm adaptively and dynamically adjust the proportion of "exploration-utilization" according to the training degree. The algorithm can capture users' dynamic interest and fully explore the action space in the recommendation system with large-scale state space. By testing the proposed algorithm model on multiple data sets,the optimal average results of *MAE* and *RMSE* are 0.16 and 0.43 respectively, which are 0.48 and 0.56 higher than the current optimal research results. Experimental results show that the proposed model is superior to the existing traditional recommendation model and recommendation model based on deep reinforcement learning.

Keywords　Recommendation system, Deep reinforcement learning, Dueling-DQN, RP-Dueling, Dynamic interest, Regret exploration

1　引言

推荐系统[1]起源于卡内基梅隆大学的罗伯特·阿姆斯特朗等,以及美国人工智能协会于1995年提出的个性化导航系统"Web,Watcher"。作为智能检索系统,推荐系统通过学习用户信息和用户历史浏览信息,为用户提供建议,帮助或引导用户向其感兴趣的内容移动。然而,由于用户兴趣的实时变化性,传统推荐算法如协同过滤方法[2]、基于内容的推荐方法[3]、基于用户兴趣的推荐方法[4]等均将数据进行静态处理,因此越来越无法满足用户和推荐平台的需求。

强化学习[5]作为机器学习[6-7]的一个分支,是一种报酬驱动的自适应学习[8]算法,它以长期收益最大化为最终目标,通过不断地迭代更新,找到一种能够使最终收益最大化的策略[9]。强化学习的这一特性弥补了传统推荐算法的不足,有效解决了传统算法无法跟踪用户兴趣变化的问题。目前,关于将强化学习与推荐系统相结合的方法的研究已有很多。例如,Rojanavasu等[10]将推荐系统与SARSA相结合,探索更多用户可能感兴趣的产品或网页,但SARSA的策略限制了算法的收敛速度,且会造成探索数据的浪费。Zheng等[11]将Q-learning应用于新闻推荐,而Q-learning算法只能应用于状态空间和动作空间较小的情况,当状态空间和动作空间较大时,算法很难学习到有效的模型。Lei等[12]将推荐系统与DQN(Human-level Control Through Deep Reinforcement Learning)相结合,根据用户的个人喜好来估计动作的价值,但当所有状态动作的价值函数不同时,状态动作价值与状态价值会有较大的偏差,这将导致模型训练效果不理想。Zhao等[13]将DDQN应用到推荐系统中,DDQN是在DQN的基础上进行改进,并没有解决状态行为值与状态值偏差较大的问题。

本文受遗憾最小化算法[14]启发,在众多研究的基础上,提出了一种基于遗憾探索的竞争网络强化学习(Dueling Net-

到稿日期:2021-06-29　返修日期:2021-10-16
通信作者:赖俊(2568754202@qq.com)

work Reinforcement Learning Based on Regret Exploration, RP-Dueling)智能推荐方法。该方法以 Dueling-DQN(Dueling Network Architectures for Deep Reinforcement Learning)为基础,在 Dueling-DQN 中加入遗憾探索机制,加速算法收敛,实现在拥有大规模状态空间的推荐系统中捕捉用户动态兴趣和对动作空间的充分探索。具体地,在每轮训练的初始阶段初始化对应各个动作的遗憾值,而后,在训练时,根据网络产生的状态值函数 $V(s)$ 和优势函数 $A(s,a)$ 的差值得到遗憾值,并根据遗憾值产生遗憾策略,最后将遗憾策略和最终的 $Q(s,a)$ 相结合,实现"探索"和"利用"的结合,并在训练过程中实现探索力度随算法收敛程度自适应动态调整。

本文将基于遗憾探索的竞争网络强化学习方法应用于推荐系统,通过学习用户历史数据,利用试错的方式估计用户兴趣转移方式。相比传统推荐方法,此方法可以很好地捕获用户动态变化的兴趣;与同类的强化学习方法相比,RP-Dueling 算法在注重用户自身的潜在兴趣在数据中的反映的同时,自适应动态调整探索力度,使得不同状态下的动作探索更充分。本文在两个数据集上离线测试了算法模型,测试结果显示,本文算法优于传统推荐算法和其他类型的强化学习推荐算法。

2 相关工作

2.1 推荐系统框架

推荐系统一般由 3 个模块组成,即用户建模模块、推荐项目建模模块和推荐算法模块。其结构示意图如图 1 所示。

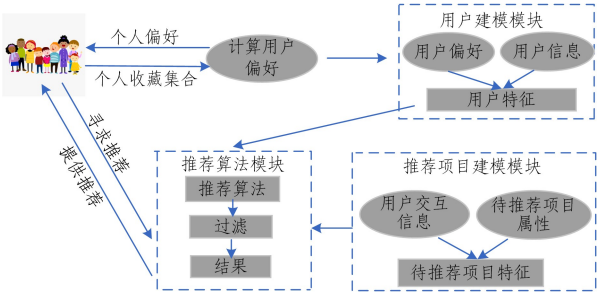


图 1 推荐系统结构框图

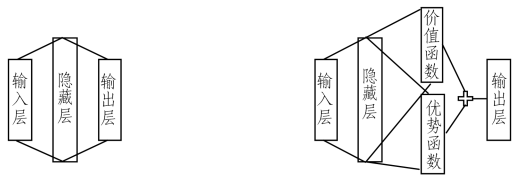
Fig. 1 Diagram of recommendation system structure

用户建模模块用于对用户进行建模,通过用户偏好和用户的个人信息,对用户进行抽象表示。例如,Yuan 等[15]使用转移学习方法对用户进行建模;Bagher 等[16]使用贝叶斯非参数模型对用户进行建模。推荐项目建模模块的作用是通过算法对用户交互序列和项目信息的学习,推断待推荐项目特征,并在推荐算法模块结合用户建模模型中的用户特征为用户进行推荐,如 Huang 等[17]使用深度神经网络搜索优质客户,He 等[18]使用非线性神经网络学习用户相似度。在推荐算法模块中,常用的算法如传统推荐算法中的基于内容的推荐[19]、协同过滤推荐[20]、基于关联规则的推荐[21]、基于效用的推荐[22]、组合推荐[23]等;以及基于深度学习的方法,如 Fu 等[24]将深度神经网络与协同过滤算法相结合,并应用于电影推荐,Li 等[25]利用 Capsule 网络来解决推荐解释问题,Gabriel 等[26]将递归神经网络应用于新闻推荐中。近年来,随着深度强化学习的快速发展,越来越多的学者将深度强化学习算法作为推荐算法,如 Chen 等[27]将强化学习应用于推荐系统中的生成对抗模型,Xiao 等[28]将深度强化学习应用于隐私感知的用户模型建模,Zhang 等[29]将基于优先经验回放的深度强化学习应用于电影推荐。

总的来说,用户建模模块通过学习用户偏好和用户信息,对用户进行建模;同时,推荐项目建模模块根据用户的交互数据和待推荐项目信息,对待推荐项目进行建模;推荐算法综合用户建模模块和推荐项目建模模块的信息,分析用户和项目的特征,使用推荐算法计算出用户可能感兴趣的产品,并根据推荐场景调整推荐结果,最后将推荐结果呈现给用户;用户收到推荐算法推荐的产品,并将对产品的操作反馈给推荐算法,推荐算法计算和分析这些反馈,并开始新一轮的推荐。

2.2 Dueling-DQN 算法简介

强化学习中环境通常被建模为马尔可夫决策过程(MDP),智能体通过与环境的互动来学习最大化预期的未来回报。智能体的行为遵循一个策略,该策略指定在 MDP 的每个状态下对可用操作的分布,智能体的目标是改进其策略以实现收益最大化。$G_t=\sum_{i=1}^{T}R_{i+1}$ 表示智能体从时刻 t 开始累积的总回报。强化学习算法以过渡元组$(s_t,a_t,r_{t+1},s_{t+1})$的形式从连续经验中学习,其中 s_t 是 t 时刻环境传给智能体的状态,a_t 是智能体在该状态下选择的动作,r_{t+1} 是智能体做出动作后,环境给予智能体的奖励,之后智能体转入下一状态 s_{t+1}。强化学习可分为两类:一类是 Value-based 算法,代表算法为 Q-learning[30];另一类是 Policy-based 算法,代表算法为 Actor-Critic[31]。本文采用了一种对抗深度强化学习算法 Dueling-DQN[32],它是一种 Value-based 算法,由 DQN[33]改进而来,与 DQN 算法不同的是,该算法中 Q 值函数的网络结构被细分为两个子结构,一个子结构生成状态值函数,另一个子结构生成优势函数。其结构对比如图 2 所示。



(a)DQN (b)Dueling-DQN

图 2 DQN 与 Dueling-DQN 的结构对比图

Fig. 2 Structure comparison of DQN and Dueling-DQN

这种结构可以表示为:

$$Q(s,a;\theta,\alpha,\beta)=V(s;\theta,\beta)+A(s,a;\theta,\alpha)$$

其中,$V(s)$是状态值函数,$A(s,a)$是优势函数,θ 表示网络结构,α,β 分别表示两个全连接层的网络参数。这种结构有两方面的优势:一方面,可以让 Q 值的学习更稳健,因为它系统地区分了哪些奖励是由状态带来的,哪些是由动作带来的,即 $V(s)$表示状态 s 的好坏,$A(s,a)$表示动作的相对平均价值的好坏。以推荐系统为例,用户的历史数据并非都可以表达用户的兴趣,有些数据仅仅是用户即兴产生的,这种数据对于用户兴趣的捕获是无用的,而这种结构可以很好地对这类数据进行区分,使这类数据对网络参数的学习产生的影响较小;另一方面,优势函数 $A(s,a)$的存在让状态 s 的所有动作 a 整体都发生了变化,因此,当同一个状态 s 下的某一个动作 a 没有

被采样到时，它的值函数 $Q(s,a)$同样也会发生改变，这很好地解决了当一个动作长期不被采样导致其采样概率越来越低的问题。同时，为解决“unidentifiable”问题，需要对优势函数进行限制，即强制优势函数估计量在选定的动作处要有零优势，可表示为：

$$Q(s,a;\theta,\alpha,\beta)=V(s;\theta,\beta)+(A(s,a;\theta,\alpha)-K)$$

其中：

$$K=\max_{a'\in|A|}A(s,a';\theta,\alpha)$$

对于任意动作 a 来说：

$$a^*=\arg\max_{a'\in A}Q(s,a';\theta,\alpha,\beta)=\arg\max_{a'\in A}A(s,a';\theta,\alpha)$$

因此可以得到：

$$Q(s,a^*;\theta,\alpha,\beta)=V(s;\theta,\beta)$$

即 $V(s;\theta,\beta)$提供了值函数的估计，$A(s,a;\theta,\alpha)$提供了优势函数的估计，在这里使用平均操作代替最大化操作，所选取的平均算子为动作空间的大小$|A|$。因此，上式可变为：

$$Q(s,a;\theta,\alpha,\beta)=V(s;\theta,\beta)+(A(s,a;\theta,\alpha)-K')$$

其中：

$$K'=\frac{1}{|A|}\sum_{a'=A}A(s,a';\theta,\alpha)$$

另外，如前所述，Dueling-DQN 算法是对 DQN 算法的改进。因此，更新网络参数时的损失函数与 DQN 相同。即：

$$L_i(\theta_i)=E_{s,a,r,s'}[(y_i-Q(s,a,\theta_i))^2]$$

其中，y_i 为目标值，其表达式为：

$$y_i=r+\gamma\max_{a'}q(s',a',\theta')$$

2.3 遗憾值及其利用方式

遗憾最小化算法用于在一般式博弈对抗中寻求近似纳什均衡，在每次做决策时，算法会计算累计遗憾值，并根据遗憾值形成动作选择的策略，以此策略进行决策。

首先，定义玩家 i 第 T 轮采样策略 σ_i 的遗憾值：

$$Regret_i^T(\sigma_i)=\sum_{t=1}^{T}(\mu_i(\sigma_i,\sigma_{-i}^t)-\mu_i(\sigma^t))$$

其中，u_i 表示玩家 i 的收益，σ^t 表示策略组，即所有玩家在 t 时刻选择动作的组合，σ_{-i}^t 表示除玩家 i 以外的其他玩家在 t 时刻选择动作的组合。通常遗憾值为负时被认为不能提升下一时刻的收益，因此当出现为负的遗憾值时，均用 0 代替，则玩家 i 在第 T 轮采取策略 σ_i 的遗憾值后，在第 $T+1$ 轮玩家 i 选择动作 a 的概率为：

$$P(a)=\frac{Regret_i^T(a)}{\sum_{b\in A}Regret_i^t(b)}$$

以两人“剪刀、石头、布”游戏为例，介绍遗憾值的含义及其应用，表 1 列出了玩家动作和收益。

表 1 玩家动作和收益对照

Table 1 Comparison of player action and income

		Play1		
		剪刀(S)	石头(R)	布(P)
Play2	剪刀(S)	0,0	−1,1	1,−1
	石头(R)	1,−1	0,0	−1,1
	布(P)	−1,1	1,−1	0,0

针对玩家 1 进行描述，若第一局玩家 1 和玩家 2 分别随机选择的动作为“石头、布”，则玩家 1 收益为−1，玩家 2 收益为 1。此时对于玩家 1，若其选择“剪刀”则收益为 1，若选择“布”则收益为 0，没有选择“剪刀”的遗憾值为：

$$\mu_1(S,P)-\mu_1(R,P)=1-(-1)=2$$

没有选择“布”的遗憾值为：

$$\mu_1(P,P)-\mu_1(R,P)=0-(-1)=1$$

则在第二局中，玩家 1 选择“剪刀、石头、布”的策略变为(2/3,0,1/3)，因此在第二局中玩家 1 更倾向于选择“剪刀”。

若在第二局中玩家 1 选择“剪刀”，玩家 2 选择“石头”，则玩家 1 每轮的遗憾值和第二轮后累加的遗憾值如表 2 所列。

表 2 累积遗憾值

Table 2 Cumulative regret value

轮次\遗憾值	剪刀(S)	石头(R)	布(P)
第一轮遗憾值	2	0	1
第二轮遗憾值	0	1	2
累积遗憾值	2	1	3

则在第三轮中玩家 1 选择“剪刀、石头、布”的策略变为(2/6,1/6,3/6)。

3 基于遗憾搜索的竞争网络架构推荐方法

RP-Dueling 原始网络结构分别对状态值函数和优势函数进行估计，减小因用户的即兴行为而产生的偏差数据对值函数学习的影响；同时遗憾探索机制可以使智能体每次决策时，会在考虑之前决策的基础上对动作空间进行充分探索。本节首先介绍 RP-Dueling 算法，之后介绍基于遗憾探索的竞争网络强化学习推荐系统框架，最后阐述模型搭建过程以及实验细节。

3.1 RP-Dueling

本文 2.3 节介绍了遗憾值及利用遗憾值产生策略的方式，定义利用遗憾值产生的策略为遗憾策略。Dueling-DQN 为 Model-free 的算法，且智能体与环境交互过程中，交互智能体每次只能得到关于一个动作的相关奖励，因此无法根据奖励来设计每个动作的遗憾值。根据本文 2.2 节对 Dueling-DQN 算法的介绍，动作优势函数 $A(s,a)$表示在状态 s 下选择动作 a 的价值，$V(s)$表示状态 s 的价值函数，且有：

$$V(s)=E_a(Q(s,a))$$

即状态的价值等于该状态下所有动作价值的期望，定义动作的遗憾值为：

$$Regret(a)=E_a(Q(s,a))-A(s,a)=V(s)-A(s,a)$$

解释为在状态 s 下不同优势值的动作在下一次交互中期待被执行的程度，即在本次交互中没有被执行的遗憾程度，且优势值越小的动作越期待被执行。之后根据遗憾值产生遗憾策略并记该遗憾策略为 RS，其表达式为：

$$RS=\left[\cdots,\frac{Regret_{a_k}}{\sum_{i=1}^{n}Regret_{a_i}},\frac{Regret_{a_{k+1}}}{\sum_{i=1}^{n}Regret_{a_i}},\cdots,\frac{Regret_{a_n}}{\sum_{i=1}^{n}Regret_{a_i}}\right]$$

其中，$Regret_{a_i}$ 表示第 k 个动作的遗憾值，$\sum_{i=1}^{n}Regret_{a_i}$ 表示所有动作的遗憾值之和。遗憾策略表示在动作选择时，根据遗憾值的大小来做决策的一种方式。本文将遗憾策略用于动作探索，同时 Dueling-DQN 中“利用”的表达式为：

$$\arg\max(Q(s,a))$$

为实现“探索”与“利用”相互作用的目的，本文更改了

原始“探索-利用”相互独立的方式，将“探索”和“利用”功能进行融合，具体方式为将 RS 与 Q 值对应位置相乘，并依据最大值策略来选择动作。即：

$$a=\arg\max(RS*Q(s,a))$$

其中，遗憾策略 RS 累积了以往决策中的经验，从全局角度考虑各动作在当前状态下的可选择性；$Q(s,a)$ 利用已学到的知识给出当前状态下各动作的得分，并利用 argmax 决策方式来选择动作，其更侧重于局部决策。公式将动作的可选择性与利用已学到的知识相结合，使得算法决策时在“利用”的同时具有“探索”性。

另外，Dueling-DQN 将 Q 函数拆分为状态值函数 V 以及优势函数 A，解决了当一个动作长期不被采样而导致其采样概率越来越低以及系统不稳定的问题。本文提出的算法 RP-Dueling 网络结构图如图 3 所示，从图中可以看出，本文并未对 Dueling-DQN 的基本结构进行更改，只是利用状态价值和优势值计算当前时刻各动作的遗憾值，该过程不涉及网络参数的计算，且在优化网络时并未改变对当前状态 s 下所有动作 a 的有关参数进行优化的目的，即并未改变 Dueling-DQN 的结构优越性。因此本文算法的设计方式并不会造成系统整体的不稳定。

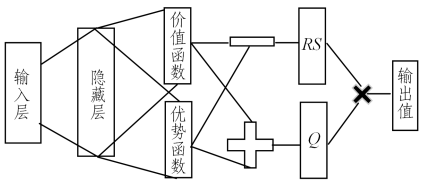


图 3 RP-Dueling

Fig. 3 RP-Dueling

RP-Dueling 中遗憾策略依赖于之前所有时刻叠加的遗憾值，该叠加方式与 2.3 节所述方式相同，是对智能体之前所有决策的统计。RP-Dueling 将遗憾策略与 Q 值函数融合，在智能体的决策过程中加入对之前决策的统计信息，使算法在“利用”已有经验的同时根据统计信息进行“探索”。在算法初始阶段，由于智能体还未学到有效经验，因此智能体更倾向于“探索”；而在算法收敛阶段，由于智能体决策带来的遗憾值逐渐减小，因此智能体更倾向于“利用”已经学到的好的经验。图 4 给出了算法训练过程中“探索”和“利用”在智能体决策中的作用随时间的变化趋势。

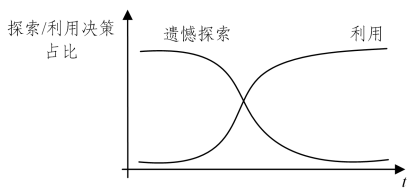


图 4 探索-利用决策占比

Fig. 4 Proportion of exploration-utilization decision

3.2 基于遗憾探索的竞争网络强化学习推荐系统框架

图 5 给出了基于遗憾探索的竞争网络强化学习推荐系统框架。

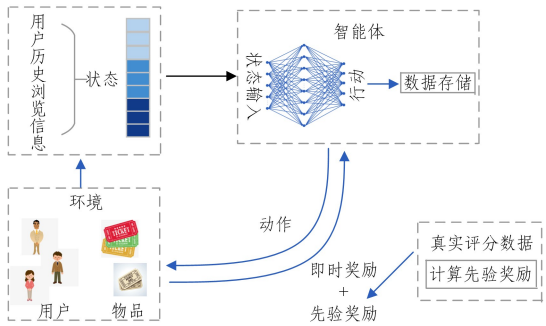


图 5 基于遗憾探索的竞争网络强化学习推荐系统框架

Fig. 5 Recommendation system framework of competitive network reinforcement learning based on regret exploration

如图 5 所示，在系统的初始阶段，环境将用户信息、电影信息和时间戳连接成为状态，然后将状态输入到智能体的网络结构中，之后网络根据状态对电影和用户进行建模，通过网络结构计算出状态值函数和融合了遗憾策略的优势函数，最终生成对应每个动作的 Q 值，并以此来选择动作。为了使智能体能更好地学习到有效经验，本文在设计奖励函数时加入了先验知识作为先验回报，在智能体做出相应决策后，先验回报计算模块计算出该状态对应的先验回报，并将先验回报加入环境给出的即时回报中形成综合奖励，环境将综合奖励值反馈给智能体，之后算法进入下一轮训练，直至训练结束。算法的伪代码如算法 1 所示。

算法 1 基于遗憾探索的竞争网络强化学习推荐系统

输入：(user_further, item_further, Times-tamp)

输出：Rate

1. 初始化 RP-Dueling 网络参数，初始化环境参数。
2. 初始化遗憾值。
3. 开始
4. For episode=1 do
5. 获取初始状态 s_t
6. For t=1, Step do
7. 根据 $a_t=\arg\max_a(Q(s,a)*RS)$ 选择动作。
8. 执行动作 a_t，从环境中获取下一状态 s_{t+1} 和奖励 r_{t+1}
9. 根据先验奖励计算公式 $e^{-abs(p\_score_{u,i}-\frac{\sum r\_score_{1,i}}{|U_{1,i}|})}$ 计算先验回报
10. 计算综合奖励 $r_{t+1}=r_{t+1}+priori\_reward$
11. 更新遗憾值
12. 将经验 $\{s_t,a_t,r_{t+1},s_{t+1}\}$ 存入经验池
13. 从经验池中获取少批量经验 $\{S_t,A_t,R_{t+1},S_{t+1}\}$
14. 计算目标值：

$$y_j=\begin{cases}r_{j+1} & \text{for terminal } s_{j+1}\\ r_{j+1}+\gamma\max_{a'}Q(s_{j+1},a';\theta') & \text{for non-terminal } s_{j+1}\end{cases}$$

15. 根据 $(y_j-Q(s_j,a_j;\theta))^2$ 利用梯度下降更新 RP-Dueling 网络
16. 更新目标网络

3.3 实验

3.3.1 实验条件

实验硬件：Intel (R) Netac SSD 120 GB+Nvidia Titan X (Pascal)+16 GB 内存。

软件环境：windows 10，TensorFlow2.5.0。

数据条件：本文使用 Movielens 中 mL-1M 和 mL-100k 数据集进行了测试，两个数据集的原始数据均被划分为 3 个独立的部分，即用户信息、电影信息和电影评分，电影评分数据集包括用户 ID、电影 ID 和用户对电影的评分。mL-1M 数据集包含 6 040 个用户对 4 900 部电影的 100 万个评分记录；mL-100k 数据集包含 943 个用户对 1 682 部电影的 10 万个评分记录。两个数据集的不同之处在于，每个用户所评分的电影数以及电影的类型复杂度不同。根据本文第三部分状态空间的定义，在数据处理过程中，以用户的评分数据表为母表，其他两个数据表依靠母表分别处理。

3.3.2 环境模型

(1)动作空间设计

物品评分预测表示通过对已知物品数据的学习，预测用户对未知物品的评分。评分等级一般为离散值，反映用户对物品的满意程度。本文实验为电影评分预测，即在给定电影信息的情况下，智能体预测用户对电影的评分。本文遵循原始数据中对电影的评分准则，将电影评级设置为 5 个等级，分别为(1,2,3,4,5)，分级值可以直接用于数值计算。在实际应用中，为了更符合环境设置，具体分数以字典的形式给出，即：

$A=\{$"$action1$":1,"$action2$":2,"$action3$":3,"$action4$":4,"$action5$":5$\}$

(2)状态空间设计

状态是智能体感知信息的载体，在训练过程中智能体通过"观察"状态中的信息来进行决策，因此状态应该包含尽可能多的信息，使智能体可以进行更全面的感知。在本文中，由环境提供给智能体的状态由用户信息(用户 id、性别、年龄、职业)和电影信息(电影 id、类型)组成。另外，为了使各个状态的信息更加丰富和完整，本文在状态栏的设计中加入了时间戳。最终设计状态空间可表示为：

$$S=(user_1\_features, movie_1\_features, timestamp_1; user_2\_features, movie_2\_features, timestamp_2; \cdots; user_n\_features, movie_n\_features, timestamp_n)$$

(3)奖励函数设计

本文在设置奖励函数时，每个状态中智能体预测的电影评分与真实电影评分无直接关联。具体设置方式为：当智能体预测的用户对电影的评分与用户对电影的真实评分一致时，环境向代理返回正奖励；当不一致时，奖励为取智能体预测的评分与用户对电影的真实评分的差值的绝对值，之后取绝对值的负值。为了给智能体一个更清晰的指导，本文在设计报酬函数时加入了先验知识。先验知识的表达式为：

$$priori=e^{-abs(p\_score_{u,i}-\frac{\sum_{l,i\in T} t\_score_{l,i}}{|U_{l,i}|})}$$

其中，abs 为取绝对值符号，$p\_score_{u,i}$ 为智能体预测的用户 u 对电影 i 的评分，$\sum_{l,i\in T} t\_score_{l,i}$ 为所有对电影 i 进行评分的用户所评的分数的总和，$|U_{l,i}|$ 为对电影 i 进行评分的用户数。表 3 列出了奖励函数的具体设计方式。

表 3 奖励函数设置

Table 3 Reward function settings

	1	2	3	4	5
1	4+$priori$	−1+$priori$	−2+$priori$	−3+$priori$	−4+$priori$
2	−1+$priori$	4+$priori$	−1+$priori$	−2+$priori$	−3+$priori$
3	−2+$priori$	−1+$priori$	4+$priori$	−1+$priori$	−2+$priori$
4	−3+$priori$	−2+$priori$	−1+$priori$	4+$priori$	−1+$priori$
5	−4+$priori$	−3+$priori$	−2+$priori$	−1+$priori$	4+$priori$

表中的第一行表示智能体预测的用户对电影的评分，第一列表示用户对电影的真实评分。

(4)数据处理

实验中，将处理后的数据按 4:1 的比例分为训练集和测试集。即利用全部数据的 80%作为智能体学习的数据，该部分数据可理解为智能体获取状态的状态空间，将剩余 20%的数据作为测试智能体学习效果的数据。在训练过程中，本文没有设置终止状态，为了符合强化学习每轮达到终止状态才结束一轮训练的要求，本文在每轮训练中设置一个固定步数，即在每轮训练中，智能体在状态空间中可以走多少步，并用计数器记录智能体已经走的步数，在每一轮开始时，计数器被重置为零。为了防止每轮训练中计数器数值溢出，本文设置了一个余数算子来限制计数器的值。此外，在每轮训练开始时，使用随机算子在状态空间中随机选择一个状态作为初始状态。本文设置每训练 10 轮进行一次测试，测试时不设最终状态，每次测试均从固定状态开始，并收集每轮测试中每步的即时奖励等数据信息。

3.3.3 实验准备

(1)评价指标

在训练过程中，本文主要观察了智能体的平均得分随训练轮数的增加而变化的趋势。平均分的计算式为：

$$\bar{R}_\tau=\frac{1}{N_\tau}\sum_{i=1}^{N_\tau} r_i$$

其中，$\bar{R}_\tau$ 表示在第 τ 轮训练中智能体所得到的平均奖励；N_τ 是超参数，表示在第 τ 轮训练中智能体所探索的步数；r_i 表示在第 τ 轮训练中第 i 步探索智能体所得到的奖励。

本文采用平均误差(MAE)和均方根误差(RMSE)作为模型评价标准[34]。这两个评价标准主要用于处理电影、歌曲等评级数据模型。表达式分别为：

$$MAE=\frac{\sum_{u,i\in T}|r_{ui}-\hat{r}_{ui}|}{|T|}$$

$$RMSE=\sqrt{\frac{\sum_{u,i\in T}(r_{ui}-\hat{r}_{ui})^2}{|T|}}$$

其中，r_{ui} 表示用户 u 对电影 i 的真实评分；$\hat{r}_{ui}$ 表示智能体预测的用户 u 对电影 i 的评分，T 表示每轮用于训练的电影数目，在本实验中其值与 N_τ 相同。

(2)实验说明

本文进行了一种算法有效性实验以及两种与其他算法的对比实验：

1)依据奖励函数设计一种平均策略,并与 RP-Dueling 算法进行实验对比,以说明本文算法的有效性;

2)与 DQN 和 DDQN 的对比实验;

3)与传统推荐算法的对比实验。

在一类实验中,我们仅收集 Reward 数据,包括平均策略收益与 RP-Dueling 算法收益,以分析 RP-Dueling 算法的有效性。

在第二类实验中,我们收集了 RMSE,MAE 和 Reward3 种数据,并通过这 3 种数据分析了各算法的效果。在第三类实验中,我们只收集了 RMSE 和 MAE 数据,并以表格的形式分析了算法的效果。

3.3.4 实验结果及分析

(1)算法有效性分析

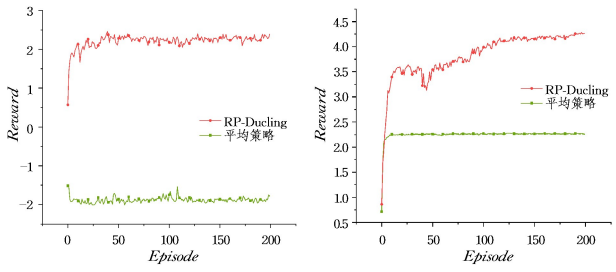
在奖励函数设置中,本文依据真实评分值对奖励进行设置,但这样会出现奖励不均衡的问题,表现为对每个评分其加和不相等,以表 3 的行和为例,具体如表 4 所列。

表 4 奖励函数和

Table 4 Sum of rewards

1	2	3	4	5
$-6+5priori$	$-3+5priori$	$-2+5priori$	$-3+5priori$	$-6+5priori$

表中第一行为评分值,第二行为评分值所对应的奖励和,可以看出,若忽略先验回报,则评分 3 所对应的奖励和最大,理论上这种不均衡会造成智能体最终偏向于将所有的电影评分都预测为 3,因为预测错误的代价是最小的。而本文提出的 RP-Dueling 算法可以很好地避免这种问题,该算法在每次决策时均会考虑之前进行决策的经验,对于真实评分不为 3 的电影,若智能体预测为 3,则会得到一个负的奖励,该负的奖励会被智能体作为每次决策的参考。同时,为证明算法的有效性,本文设计了一种平均策略算法,即在训练过程中每次决策时固定预测评分均为 3,以此来训练模型,并利用该模型进行测试。在两个数据集上进行测试,并将平均策略测试结果与 RP-Dueling 算法测试结果进行对比,结果如图 6 所示。



(a)在 mL-1M 数据集上的测试结果 (b)在 mL-100k 数据集上的测试结果

图 6 固定评分与依据策略评分误差棒图(电子版为彩图)

Fig. 6 Error bar chart of fixed score and score according to strategy

图 6 中红色曲线均为由 RP-Dueling 算法依据策略进行预测评分的平均累积回报值曲线,而绿色曲线均为利用平均策略进行预测评分的回报值曲线。从图 6(a)和图 6(b)中可以看出,依据 RP-Dueling 进行评分的平均累积回报远优于依据平均策略进行评分的回报。证明了在奖励函数不均衡的条件下,本算法在决策时并不会偏向于选择评分 3 这个动作,从而证明了 RP-Dueling 算法的有效性。

(2)与深度强化学习算法进行实验对比与分析

为了证明本文算法更具说服力,在两个数据集上进行实验时,在保持训练轮数、每轮步数、折扣因子等参数不变的情况下,只将基本算法由 RP-Dueling 改为 DQN 和 DDQN,并进行比较。实验结果如图 7 和图 8 所示。

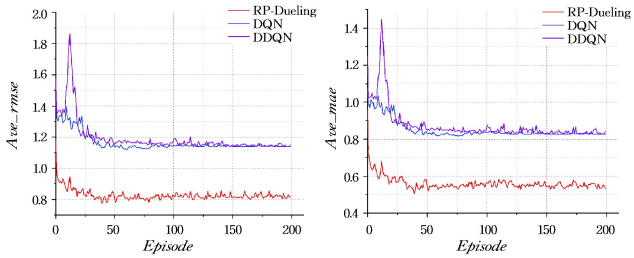


图 7 mL-1M 实验结果

Fig. 7 Results on mL-1M

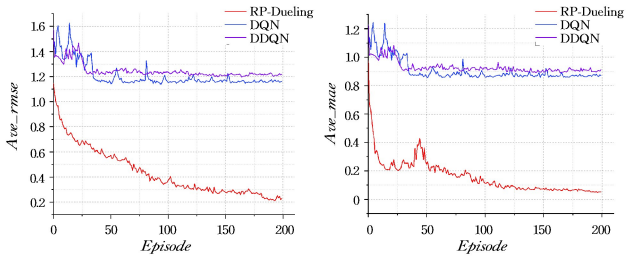


图 8 mL-100k 实验结果

Fig. 8 Results on mL-100k

考虑到测试数据量庞大,因此本文采用平均 RMSE(Ave_rmse)和平均 MAE(Ave_mae)对算法进行评估。从图 7 和图 8 可以看出,随着训练轮数的增加,Ave_mae 和 Ave_rmse 都逐渐减小。在这两个数据集上,本文方法优于其他两种方法;同时,我们发现在 DQN 和 DDQN 算法的实验中,Ave_mae 和 Ave_rmse 曲线相似。经过分析,我们认为这种相似结果主要来自 DDQN 和 DQN 的相似性。DDQN 由 DQN 改进而来,这种改进主要是为了解决 DQN 高估 Q 值的问题。这两种算法的更新式如下,DQN 的更新式为:

$$Q(s,a|\theta)\leftarrow Q(s,a|\theta)+L$$

其中:

$$L=\alpha[r+\gamma\max_{a'}\hat{Q}(s',a'|\theta^{-})-Q(s,a|\theta)]$$

DDQN 的更新式为:

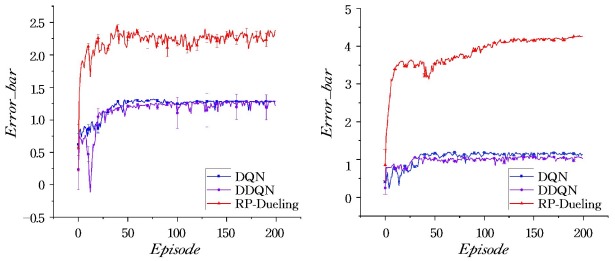
$$Q(s,a|\theta)\leftarrow Q(s,a|\theta)+L'$$

其中:

$$L'=\alpha[r+\gamma\max_{a'}\hat{Q}(s',\max_{a'}Q(s',a'|\theta)|\theta')-Q(s,a|\theta)]$$

这两个公式只是在 Q 值的更新方式上有所不同,而两种算法的网络结构是相同的,这种差异不会造成结果上的巨大差异。因此,我们有理由相信,两种算法的实验结果相似为正常现象。

误差棒是检验实验可靠性以及算法可复现性的方法之一。它以测量值的算术平均值为中点,在指示测量值大小的方向上画一条线,线的一半长度等于不确定度。测线长度越短,实验结果越可靠。图 9 中给出了 3 种算法平均奖励的误差棒曲线。



(a)mL-1M 数据集 (b)mL-100k 数据集

图 9 mL-1M 和 mL-100k 数据集上的 Dueling-DQN，DQN，DDQN 的误差棒

Fig. 9 Error bar of Dueling-DQN，DQN，DDQN on mL-1M and mL-100k datasets

如图 9 所示，在两个数据集上的实验结果表明，RP-Dueling 在结果上的不确定性明显小于其他两类算法，且在算法收敛阶段，在最大不确定性情况下，本文算法的效果仍优于其他两种算法。图 9 中的细节如表 5 所列。

表 5 3 种算法在 Max_reward，Min_reward，Average_reward，Standard_deviation，Average_reward_add 上的结果

Table 5 Results of three algorithms on Max_reward，Min_reward，Average_reward，Standard_deviation and Average_reward_add

Datasets	Method Evaluation	Max_ reward	Min_ reward	Average_ reward	Standard_ deviation	Average_ reward_add Compare
mL-1M	RP-Dueling	**2.39**	0.57	**2.21**	0.0690	—
	DDQN	1.24	−0.11	1.15	0.0511	+1.15
	DQN	1.32	**0.62**	1.25	**0.0267**	+1.07
mL-100k	RP-Dueling	**4.28**	**0.85**	**3.83**	**0.0551**	—
	DDQN	1.09	0.25	0.99	0.1016	+3.19
	DQN	1.19	0.23	1.06	0.1947	+3.09

表 5 中共收集了两个数据集中 4 个评测指标的数据(Max_reward，Min_reward，Average_reward，Standard_deviation)和一个对比结果数据(Average_ reward_add)，并将表现较好的评测指标结果突出显示。经观察发现，在 mL-100k 数据集中，RP-Dueling 算法在 4 个评测指标中的测试结果均远优于另外两种算法；而在 mL-1M 数据集中，RP-Dueling 在 Max_reward 和 Average_reward 上的测试结果仍然优于其他两种算法，但是在 Min_reward 和 Standard_deviation 的测试结果劣于另外两种算法，分析其原因为：在 mL-1M 数据集中，每个用户的历史数据都很庞大，这导致算法对用户兴趣的提取出现了较大的方差，导致最终结果相比其他两种算法震荡幅度较大。但是这种震荡并不会对最终的结果产生过大的影响。综合图 9 和表 5 可以证明本文提出的算法具有高可靠性和很好的可复现性。

(3)与传统推荐算法进行实验对比与分析

在工业领域，虽然目前基于深度强化学习和基于深度学习的推荐算法得到越来越多的应用，但是起主导作用的仍然是传统推荐算法，因此本文选取的用作对比的传统推荐算法均是过去或现在具有很大影响力的算法，包括：

1)KNNBasic[35]。该算法用于解决未知物品 u 的评价问题，其只需要找到 K 个与 u 相似的已知物品，之后通过 K 个已知物品再对 u 进行评估。

2)KNNWithMeans[36]。该算法的基本假设为用户对物品的评分有高有低，考虑每个用户对物品评分的均值或者每个物品得分的均值，去除参考用户打分整体偏高或者偏低的影响。

3)KNNBaseline[37]。该算法考虑了计算的偏差，偏差的计算基于 Baseline。

4)SVD[38]。该方法首先将用户的评分矩阵分解为用户隐向量矩阵和物品隐向量，在预测用户对某一物品的评分时直接用用户隐向量与物品隐向量进行内积运算，即可得出用户对物品的预测评分。可以表示为：

$$\hat{r}_{ui}=p_u\boldsymbol{q}_i^{\mathrm{T}}=\sum_{k=1}^{K}p_{uk}\boldsymbol{q}_{ki}^{\mathrm{T}}=\sum_{k=1}^{K}p_{uk}\boldsymbol{q}_{ik}\approx r_{ui}$$

5)SVD++[39]。该算法是在 SVD 模型的基础上加入了用户对物品的隐式行为。此时可以认为评分＝显式兴趣＋隐式兴趣＋用户对物品的偏见。可以表示为：

$$\hat{r}_{ui}=\mu+b_i+b_u+(p_u+|N(u)|^{-0.5}\sum_{i\in N(u)}y_i)\boldsymbol{q}_i^{\mathrm{T}}$$

6)NMF[40]。该算法的思想是，对于任意给定的非负矩阵 $\boldsymbol{A}$，NMF 算法可以求出一个非负矩阵 $\boldsymbol{U}$ 和一个非负矩阵 $\boldsymbol{V}$，即将一个非负矩阵分解为两个非负矩阵的乘法形式。该算法在用于推荐时，已知用户信息矩阵、物品信息矩阵以及用户对物品的评分，利用该算法可以预测用户对其他未知物品的评分。

该对比实验主要收集了在两个数据集上的 RMSE 和 MAE 结果，实验数据如表 6 所列，并将最佳结果进行了突出显示。可以看出，RP-Dueling 在两个数据集上的测试结果均优于传统推荐算法，这也证明了 RP-Dueling 应用于推荐系统比一般推荐算法更有效。

表 6 RP-Dueling 与传统算法在 *RMSE* 和 *MAE* 上的测试结果

Table 6 *RMSE* and *MAE* comparison of RP-Dueling and traditional algorithms

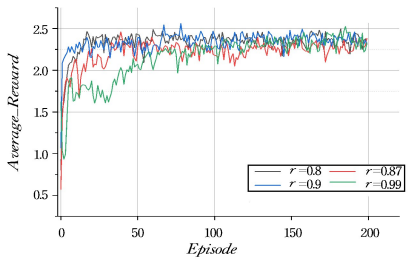
Datasets	Evaluation Method	RP-Dueling	KNNBasic	KNNWithMeans	KNNBaseline	SVD	SVD++	NMF
mL-1M	*RMSE*	**0.8200**	0.9227	0.9290	0.8949	0.8730	0.8611	0.9167
	MAE	**0.5500**	0.7273	0.7383	0.7064	0.6858	0.6717	0.7246
mL-100k	*RMSE*	**0.4300**	0.9790	0.9511	0.9294	0.9349	0.9195	0.9638
	MAE	**0.1600**	0.7727	0.7493	0.7321	0.7365	0.7212	0.7583

3.3.5 超参数分析

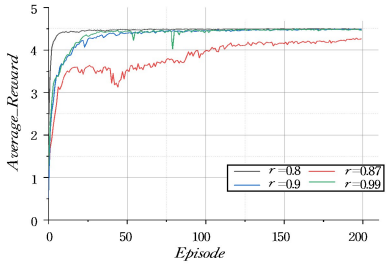
本节对折扣因子 γ 的影响进行分析。折扣因子是马尔可夫决策过程的一部分，它通过指数方案来估计未来报酬，其表达式为：

$$G_t=r_t+\gamma r_{t+1}+\gamma^2 r_{t+2}+\cdots+\gamma^n r_{t+n}$$

该等式保证了 Bellman 方程[41]的理论收敛性，其有效性在文献[42]中得到了证实。本节通过实验来探究不同 γ 值对结果的影响，实验中其他参数值保持不变。实验结果如图 10 所示。



(a)mL-1M



(b)mL-100k

图 10 不同折扣因子下回报趋势图

Fig. 10 Growth trend of rewards at different discount factors

本文在[0.8,1]范围内选择了 4 个不同的折扣因子值,并在两个数据集上进行了测试。在 mL-1M 数据集上的结果表明,在不同的折扣因子下,模型的最终收益率变化并不明显。在 mL-100k 数据集上进行测试,在相同轮数的训练下,当折扣因子值为 0.87 时,模型的收敛速度明显慢于其他 3 条曲线,但不影响最终的收敛。在两个数据集上的实验结果表明,不同的折扣因子在不同的数据集上对实验结果均有影响,且影响方式不规律,但均不影响最终算法的收敛。

结束语 本文在两个公共数据集上测试了 RP-Dueling 算法和对比算法。实验结果表明,RP-Dueling 算法在实验效果和重现性方面均优于其他同类算法和传统算法。同时,本文对模型中的折扣因子进行了实验分析,实验结果表明该超参数不同的值对算法最终的收敛影响较小。此外,本文还分析了 RP-Dueling 算法能够产生更好结果的原因:

(1)与传统算法相比,RP-Dueling 算法是一种自适应算法,能够通过试错学习自动总结模型,减少人为干预,提高算法的可信度,具有实时性。这两个性质是传统算法所不具备的,对于推荐算法本身来说,这两个性质更符合推荐的要求。

(2)RP-Dueling 算法不直接生成 Q 值,而是输出优势值和状态值,通过这两个值的组合生成 Q 值和遗憾策略,动作选择由最大 Q 值变为最大 Q 值与遗憾策略的乘积。由于遗憾策略包含以往的信息,因此智能体在做决策时会综合当前信息和以往的信息,这加速了算法的收敛,提高了智能体决策的准确性。

参 考 文 献

[1] JACOBI J A, BENSON E A, LINDEN G D. Recommendation system: U. S. Patent 7,908,183[P]. [2011-3-15]. https://patents.glgoo.top/patent/US7908183B2/en.

[2] SCHAFER J B, FRANKOWSKI D, HERLOCKER J, et al. Collaborative filtering recommender systems[M]// The Adaptive Web. Berlin: Springer Press, 2007: 291-324.

[3] DORSCH M, QIU Y, SOLER D, et al. PK1/EG-VEGF induces monocyte differentiation and activation[J]. Journal of Leukocyte Biology, 2005, 78(2): 426-434.

[4] QI H M, LIU Q, DAI D X. Personalized Friend Recommendation based on Interest Topics [J]. Computer Engineering and Science, 2018, 40(2): 348-353.

[5] SUTTON R S, BARTO A G. Reinforcement learning: An introduction[M]. USA: MIT Press, 2018.

[6] MOHRI M, ROSTAMIZADEH A, TALWALKAR A. Foundations of machine learning[M]. USA: MIT Press, 2018.

[7] JORDAN M I, MITCHELL T M. Machine learning: Trends, perspectives, and prospects[J]. Science, 2015, 349(6245): 255-260.

[8] MESSNER W, HOROWITZ R, KAO W W, et al. A new adaptive learning rule[C]// Proceedings of IEEE International Conference on Robotics and Automation. New York: IEEE Press, 1990: 1522-1527.

[9] KAELBLING L P, LITTMAN M L, MOORE A W. Reinforcement learning: A survey[J]. Journal of Artificial Intelligence Research, 1996, 4(1): 237-285.

[10] ROJANAVASU P, SRINIL P, PINNGERN O. New Recommendation System Using Reinforcement Learning[J]. International Journal of the Computer, the Internet and Management, 2005, 13(3): 23.

[11] ZHENG G, ZHANG F, ZHENG Z, et al. DRN: A deep reinforcement learning framework for news recommendation[C]// 27th International World Wide Web(WWW 2018). Association for Computing Machinery, 2018: 167-176.

[12] LEI Y, WANG Z, LI W, et al. Social attentive deep q-network for recommendation[C]// Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval. 2019: 1189-1192.

[13] ZHAO Z, CHEN X. Deep Reinforcement Learning based Recommend System using stratified sampling[C]// IOP Conference Series: Materials Science and Engineering. IOP Publishing, 2018.

[14] ZINKEVICH M, JOHANSON M, BOWLING M, et al. Regret minimization in games with incomplete information[J]. Advances in Neural Information Processing Systems, 2007, 20(14): 1729-1736.

[15] YUAN F, HE X, KARATZOGLOU A, et al. Parameter-efficient transfer from sequential behaviors for user modeling and recommendation[C]// Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. 2020: 1469-1478.

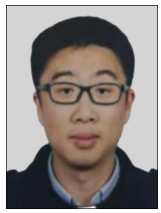
[16] BAGHER R C, HASSANPOUR H, MASHAYEKHI H. User trends modeling for a content-based recommender system[J]. Expert Systems with Applications, 2017, 87: 209-219.

[17] HUANG Z, SHAN G, CHENG J, et al. TRec: An efficient recommendation system for hunting passengers with deep neural networks[J]. Neural Computing and Applications, 2019, 31(1): 209-222.

[18] HE X, HE Z, SONG J, et al. Nais: Neural attentive item similarity model for recommendation [J]. IEEE Transactions on

Knowledge and Data Engineering,2018,30(12):2354-2366.
[19] PAZZANI M J,BILLSUS D. Content-based recommendation systems[M]// The Adaptive Web. Berlin:Springer Press,2007:325-341.
[20] BREESE J S,HECKERMAN D,KADIE C. Empirical Analysis of Predictive Algorithms for Collaborative Filtering[J]. Uncertainty in Artificial Intelligence,2013,98(7):43-52.
[21] LIN W,ALVAREZ S A,RUIZ C. Efficient Adaptive-Support Association Rule Mining for Recommender Systems[J]. Data Mining & Knowledge Discovery,2002,6(1):83-105.
[22] YIN Y,FENG D,SHI S. A Utility based personalized article recommendation method[J]. Journal of Computer Science,2017,40(12):2797-2811.
[23] VARTAK M,MADDEN S. CHIC:a combination-based recommendation system[C]// Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data. 2013:981-984.
[24] FU M,QU H,YI Z,et al. A novel deep learning-based collaborative filtering model for recommendation system[J]. IEEE transactions on cybernetics,2018,49(3):1084-1096.
[25] LI C,QUAN C,PENG L,et al. A capsule network for recommendation and explaining what you like and dislike[C]// Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval. 2019:275-284.
[26] GABRIEL DE SOUZA P M,JANNACH D,DA CUNHA A M. Contextual hybrid session-based news recommendation with recurrent neural networks[J]. IEEE Access,2019,7:169185-169203.
[27] CHEN X,LI S,LI H,et al. Generative adversarial user model for reinforcement learning based recommendation system[C]// International Conference on Machine Learning. PMLR,2019:1052-1061.
[28] XIAO Y,XIAO L,LU X Z,et al. Deep Reinforcement Learning-Based User Profile Perturbation for Privacy Aware Recommendation[J]. IEEE Internet of Things Journal,2020,8(6):4560-4568.
[29] ZHANG Y Y,SU X Y,LIU Y. A Novel Movie Recommendation System Based on Deep Reinforcement Learning with Prioritized Experience Replay[C]// 2019 IEEE 19th International Conference on Communication Technology (ICCT). New York:IEEE,2019:1496-1500.
[30] WATKINS C J C H,DAYAN P. Q-learning [J]. Machine learning,1992,8(3/4):279-292.
[31] PETERS J,SCHAAL S. Natural Actor-Critic [J]. Neurocomputing,2008,71(7/8/9):1180-1190.
[32] WANG Z,SCHAUL T,HESSEL M,et al. Dueling network architectures for deep reinforcement learning[C]// International Conference on Machine Learning. PMLR,2016:1995-2003.
[33] FAN J,WANG Z,XIE Y,et al. A theoretical analysis of deep Q-learning[C]// Learning for Dynamics and Control. PMLR,2020:486-489.
[34] XIANG L. Recommended system practice[M]. Beijing:Posts & Telecom Press. 2012.
[35] HERLOCKER J L,KONSTAN J A,TERVEEN L G,et al. Evaluating collaborative filtering recommender systems[J]. ACM Transactions onInformation Systems(TOIS),2004,22(1):5-53.
[36] COLLINS A,TKACZYK D,BEEL J. A Novel Approach to Recommendation Algorithm Selection using Meta-Learning [C]// AICS. 2018:210-219.
[37] YANG K X,LI Y W. Development and Design of mobile Intelligent Learning Platform based on Collaborative Filtering Algorithm [J]. Software Engineering and Applications,2019,8(3):104-114.
[38] AHARON M,ELAD M,BRUCKSTEIN A. K-SVD:An algorithm for designing overcomplete dictionaries for sparse representation[J]. IEEE Transactions on Signal Processing,2006,54(11):4311-4322.
[39] KOREN Y. Factorization meets the neighborhood:a multifaceted collaborative filtering model[C]// Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2008:426-434.
[40] WANG X,YANG H,LIM K. Privacy-preserving POI recommendation using nonnegative matrix factorization [C]// 2018 IEEE Symposium on Privacy-aware Computing(PAC). New York:IEEE,2018:117-118.
[41] BARRON E N,ISHII H. The Bellman equation for minimizing the maximum cost[J]. Nonlinear Analysis:Theory,Methods & Applications,1989,13(9):1067-1090.
[42] AMIT R,MEIR R,CIOSEK K. Discount factor as a regularizer in reinforcement learning[C]// International Conference on Machine Learning. PMLR,2020:269-278.

HONG Zhi-li, born in 1994, postgraduate. His main research interests include deep reinforcement learning, recommendation system and game confrontation.



LAI Jun, born in 1979, postgraduate, associate professor, master supervisor. His main research interests include deep reinforcement learning and command information system engineering.

(责任编辑:喻藜)